

Application of a Quantum Ensemble Model to Linguistic Analysis

Andrij Rovenchak^{1*} and Solomija Buk²

¹ Department for Theoretical Physics,

² Department for General Linguistics,

Ivan Franko National University of Lviv, Ukraine

November 24, 2010

Abstract

A new set of parameters to describe the word frequency behavior of texts is proposed. The analogy between the word frequency distribution and the Bose-distribution is suggested and the notion of “temperature” is introduced for this case. The calculations are made for English, Ukrainian, and the Guinean Maninka languages. The correlation between in-deep language structure (the level of analyticity) and the defined parameters is shown to exist.

Keywords: Word frequency; Text parameters; Bose-distribution

*Corresponding author: A. Rovenchak, Department for Theoretical Physics, 12 Drahomanov St., Lviv, UA-79005, Ukraine; tel.: +380 32 2614443, e-mail: andrij.rovenchak@gmail.com, andrij@ktf.franko.lviv.ua

1 Introduction

Quantitative analysis of large text samples revealed regularities in the behavior of various text parameters. The empirical laws found in texts, such as Zipf’s law, are known to hold in various domains, in particular the distribution of nucleotides in genomes and other fields of biology [1, 2, 3, 4], regularities in social sciences [5, 6, 7, 8, 9], etc.

Approaches from the domain of statistical physics can be used to study systems composed of many units in general, and texts are suitable for such studies as well. The application of physical techniques in linguistic is quite common [10, 11, 12, 13, 14], other domains are also successfully covered by physical approaches, cf. [15].

In this work, we analyze quantitative behavior of texts by finding analogy with a bosonic system within grand canonical ensemble. In doing so, we demonstrate the possibility to assign some new parameters characterizing the frequency structure of texts, one of which can be conventionally called “temperature”.

The notion of “temperature of texts” was discussed from different points of view by several authors. Mandelbrot [16] suggested the name “informational temperature of texts” for a parameter in a rank–frequency distribution (known as the Zipf–Mandelbrot law). Such a parameter is related to “good” or “bad” employment of words, especially rare words [17]. The “temperature” as a measure of communicative ability was introduced in [18]. Recently, Miyazima and Yamamoto [19] used the classical Boltzmann distribution to define the “temperature of texts” from the frequency data of the most frequent words. We propose a different approach, mainly addressing the behavior of low-frequency vocabulary.

The paper is organized as follows. In Section 2 we recall main notions used in further text, namely the principles of rank–frequency distribution compilation as well as the term *hapax legomena*. Section 3 contains main part, where the physical analogy with the Bose-distribution is discussed in detail and parameters of text frequency distribution are given suitable interpretation. The results of text analysis in three languages are given in Section 4, and Section 5 contains brief discussion of the presented approach.

2 Rank–frequency distribution

In this work, we analyze texts on the word level. While the notion of “word” has no unique definition, cf. [20], we restrict ourselves to the so called “orthographic word” defined as an alphanumeric sequence between two spaces or punctuation marks. Different word forms, like ‘hand’ and ‘hands’, ‘write’ and ‘wrote’, etc. are considered as different words for simplicity.

To obtain a rank–frequency distribution, one should first compile the frequency list from a given sample. Then, the item with the highest frequency is given **rank 1**, the second most frequent item is given **rank 2**, and so on. The items with the same frequency are given a consecutive range of ranks, the ordering within which can be arbitrary.

The studies of rank–frequency distributions originate from text analysis, and despite the regularities found there are known to hold in various domains, texts still remain the most easily accessible material having a good variety of sorts to be analyzed.

A typical rank–frequency distribution has the shape shown in Fig. 1.

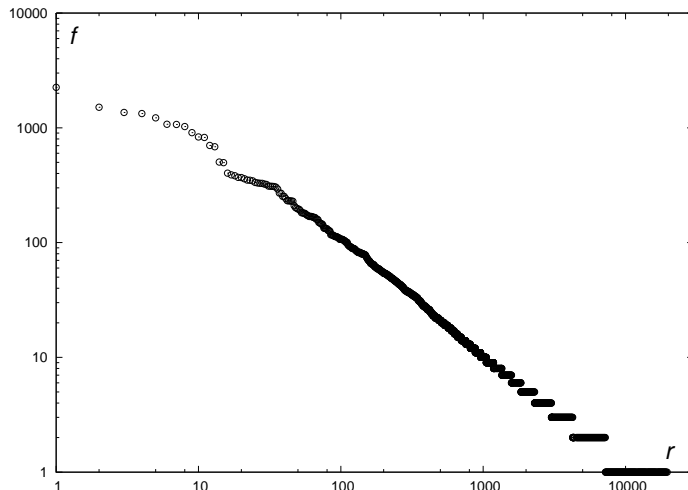


Figure 1: Typical rank–frequency distribution. The absolute frequency f is shown versus the rank r for orthographic words of *Perekhresni stežky* [*The Cross-Paths*], a Ukrainian novel by Ivan Franko. Data are obtained by the authors on the preliminary stage of compiling the frequency dictionary of the novel [21].

Horizontal plateaus in the domain of high ranks / low frequencies correspond to a large number of words having the same frequency. The longest plateau correspond to frequency 1. Such words are known as *hapax legomena*, the term originating from Bible studies.

Hapax legomena is a Plural of the Classical Greek term *hapax legomenon* (ἁπαξ λεγόμενον) translated as ‘[something] said [only] once’. That is, this term corresponds to the tokens appearing only once in a given sample. Examples from Bible include [22]: לִילִית ‘Lilith’ (a word of obscure meaning) or עֵצִי-גִפְרִי ‘gopher wood’ (used to build Noah’s Ark). Other often cited examples are: αὐτόγυον, a kind of plough (Hesiod, Ἔργα καὶ Ἡμέραι [*Opera et Dies = Works and Days*], 433); *honorificabilitudinitatibus* ‘the state of being able to achieve honors’ (Shakespeare, *Loves Labours Lost*, act 5, scene 1 [23, p. 372]).

For large text samples, about 40 to 60 per cent of occurring words are hapaxes, depending on the text size [24, p. 72]. The relative number of *hapax legomena* slightly decreases as the text becomes longer. Various quantities depending on the text size N are well described by the power law [25], and the number of hapaxes fits into this family as well,

$$N_{\text{hapax}} \stackrel{?}{=} AN^b.$$

Note, however, that for statistical studies texts must be sufficiently long.

Indeed, even in such a long sentence having twenty-three tokens all the words are hapaxes, except for “hapaxes” themselves since they occur twice.

3 Physical analogy

The rank–frequency distribution of words in texts has clear similarities with Bose-distribution in statistical physics. We suggest to identify the energy level numbers j with word frequencies (the number of occurrences in a given text). Thus, the words with frequency 1 occupy the level $j = 1$, the words with frequency 2 occupy the level $j = 2$, etc. The level occupation then corresponds to the number of different words with the same frequency. Since the level occupation can reach any value (in particular, significantly larger than unity) the use of the Bose-distribution is appropriate. The lowest level corresponds to *hapax legomena* and in this scheme can be identified with the Bose-condensate.

3.1 Defining energy spectrum

In the Bose-distribution the occupation of the j th level is given by

$$N_j = \frac{1}{z^{-1}e^{\varepsilon_j/T} - 1}, \quad (1)$$

where z is the fugacity, ε_j is the energy of the j th level, and T is the temperature.

As shown further, a power energy spectrum gives a proper description for lower levels,

$$\varepsilon_j = (j - 1)^\alpha. \quad (2)$$

The unity is subtracted to ensure that the lowermost level has zero energy.

Due to the nature of the frequency distribution, a simple model of a very weak log-of-log growth is appropriate for the energy spectrum at high levels, $\varepsilon_j \propto \ln \ln j$ for $j \gg 1$, cf. Fig. 2. Note, however, that a log-of-log spectrum requires the maximal number of levels to be bounded from above by some j_{\max} .

3.2 Parameters of the Bose-distribution

We defined the parameters in Eq. (1) in two steps. First, the parameter z , being interpreted as fugacity in physics, is defined from the occupation number of the lowermost state, i. e., the number of *hapax legomena*:

$$N_{\text{hapax}} = \frac{z}{1 - z}. \quad (3)$$

“Temperature” T and exponent α in Eq. (2) are found simultaneously by fitting the occupation of higher energy levels to

$$N_j = \frac{1}{z^{-1}e^{(j-1)^\alpha/T} - 1} \quad (4)$$

via two parameters, α and T . The sample results of fitting are presented in Fig. 2. These calculations, as well as other given further in these work, were made using the nonlinear least-squares Marquardt–Levenberg algorithm implemented in the `fit` procedure of GnuPlot, version 4.0.

One should note that the parameter T is dimensionless in our case, as is the energy ε_j . Such definition differs, e. g., from [19], where a distribution of some standard text was used to set the reference temperature in Kelvins.

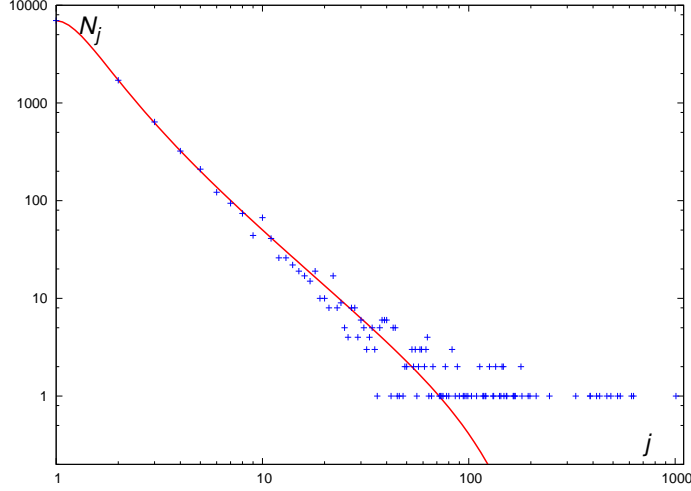


Figure 2: (Color online) The fit of the power energy spectrum to the level occupations. Blue crosses correspond to the data obtained by the authors on the basis of first 40 chapters (of total 60) from the text mentioned in the caption of Fig. 1. Solid line is the fitting curve (4) for the first 20 values of occupation numbers N_j .

The state with $T = 0$ corresponds to all the frequencies equal to unity, that is, the whole text is composed of *hapax legomena*. This could be the case of a very short text, not longer than just one or a couple of sentences (cf. the example at the end of Section 2).

Presently, we fit first 10–20 levels using the power excitation spectrum (2). Higher levels are neglected since a different dependence on j must be applied to ensure good fitting of the occupation data N_j , a suggested in the previous subsection. The parameter T obtained in such a way scales (very precisely) as N^β ($\beta < 1$). The scaling is related to the definition of “thermodynamic limit” for the problem under consideration. Just to recall, in the system of N bosons trapped to a D -dimensional harmonic oscillator potential with frequency ω the thermodynamic limit is given by $\omega N^{1/D} = \text{const}$ as $N \rightarrow \infty, \omega \rightarrow 0$ [27]. Since ω (or $\hbar\omega$ if Planck’s constant \hbar is not set equal to unity) is a natural unit for the oscillator energy, the power-like scaling of the quantities measured in the energy units is expectable for the and for the systems with power energy spectrum as well.

Curiously, the ratio $\ln T / \ln N$ exhibits an insignificant variation with the

size of the text sample (for a sufficiently long text). This makes it a good variable for comparative linguistic studies.

4 Some results

So far, we have performed analysis of some texts written in English (Germanic language), Ukrainian (Slavic language), as well as Guinean Maninka (in the Nko script; a language from the Mande family). Such a vast choice is suggested to check the approach on significantly different language materials in order to reveal both universal and unique features of the parameter behavior.

Fig. 3 demonstrates the “temperature” behavior of an English text (*Moby-Dick* by Hermann Melville) and two novels in Ukrainian (*Perekhresni stežky* [*The Cross-Paths*] by Ivan Franko and *Sobor* [*The Cathedral*] by Oles Honchar).

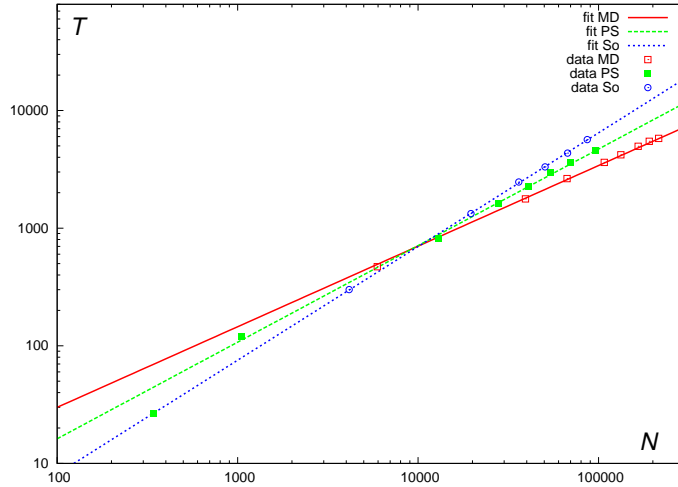


Figure 3: (Color online) The behavior of “temperature” as the size of text grows. MD — *Moby-Dick*; PS — *Perekhresni stežky*; So — *Sobor*). The lines correspond to the linear fits of the data represented by the respective symbols.

Table 1 shows the numerical data on the parameter T calculated by grasping increasing shares of chapters. The data based on an article from a

Guinean journal *Yélen'* are also given.

The values of z in all the cases are close to 1. A better resolution might be achieved by introducing an analog of the chemical potential μ in a standard way, $z = e^{\mu/T}$.

The fitting gives the values of the exponent α slightly decreasing as the text size grows. An interpretation in terms of an external potential can be applied to justify such a change. Indeed, if the presence of an external potential is treated in the semiclassical approach [26], the decreasing values of the exponent in a power excitation spectrum effectively correspond to weakening of the steepness of an external potential. That is, as a text becomes longer, it suffers less from some external influences.

Indeed, in one dimension a power energy spectrum $\varepsilon_p \propto p^\alpha$ leads to the density of states

$$g(\varepsilon) \propto \varepsilon^{\frac{1}{\alpha}-1}. \quad (5)$$

On the other hand, non-interacting particles confined into trapping potential $U(x) \propto x^\eta$ in the semi-classical approach [26] have the density of states

$$g(\varepsilon) \propto \varepsilon^{\frac{1}{\eta}-\frac{1}{2}}. \quad (6)$$

Note that a rigid box corresponds to $\eta = \infty$.

Thus, an effectively occurring exponent α is related to η via

$$\alpha = \frac{2\eta}{\eta + 2} \quad (7)$$

leading to $\alpha = 3/2$ for $\eta = 6$ and $\alpha = 1$ for $\eta = 2$.

Preliminary, the obtained T and α -exponent values correlate with the analyticity level of the language. Lower values correspond to higher analyticity (less word inflection), as can be seen from the opposition between English and Ukrainian (both Indo-European languages). So far, we do not have sufficient data to make further statements, in particular, for the language from an unrelated language family (Mande), which data are given for curiosity and future references. As should be expected, a low value of α for the Maninka sample suggests a high level of analyticity.

Finally, in Fig. 4 we present the results of “temperature” calculation made for short Ukrainian texts of different genres [28]. Close values denote weak genre dependence of this parameter. A multivariate discriminant analysis is required to study this issue in more detail.

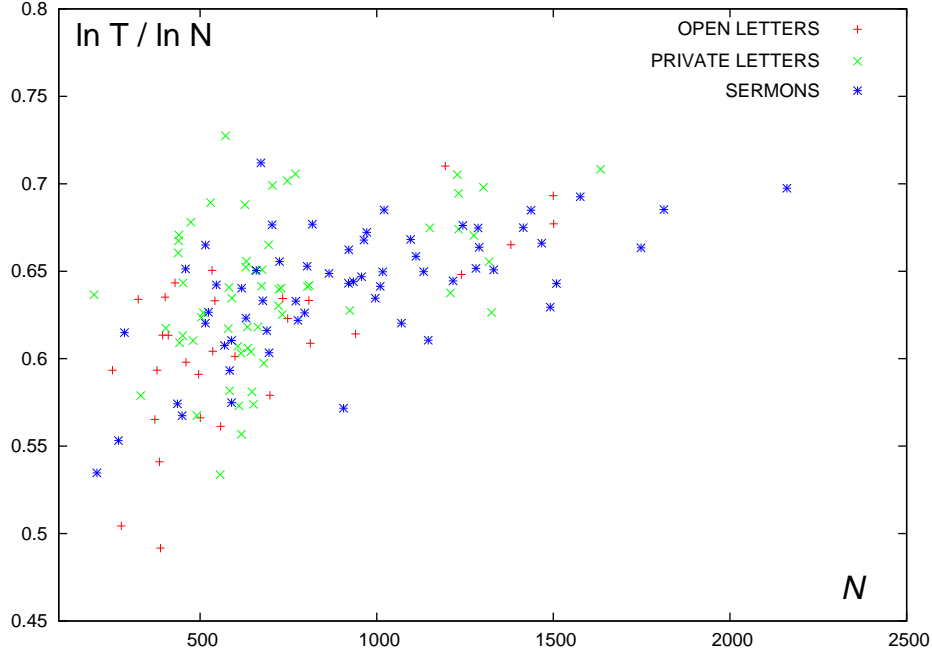


Figure 4: (Color online) The behavior of $\ln T / \ln N$ for texts from different genres. Open letters, private letters, and sermons are shown.

5 Brief discussion

The presented results are a preliminary attempt to define a new set of parameters describing frequency structure of texts. Further application of this approach to a larger number of texts from different languages written by different authors is required to establish the correlation between parameter values and text language/authorship. The shape of the spectrum in the whole domain of the variation of j must be considered in further studies to give a proper description of the level occupation. Also, more parameters can be calculated within the “thermodynamic approach” (like some analogs of total energy, specific heat, etc., cf. [18]). One of the tasks which we expect from such calculations is the possibility of automatic text attribution useful for automated language processing. Applications beyond linguistics – in genetics, social sciences, etc. – are also possible in future.

Acknowledgements

We are grateful to anonymous Referees for useful critical comments and suggestions, which were helpful in improving the manuscript.

This work was partly supported by Project No. M/6-2009 from the Ministry of Education and Sciences of Ukraine and WTZ Project UA 05/2009 from the Österreichischer Austauschdienst.

References

- [1] M. L. Bender and Pr. Gill, *Current Anthropology* **27**, 280-283 (1986).
- [2] Neng-zhi Jin, Zi-xian Liu and Wen-yuan Qiu, *Chin. J. Chem. Phys.* **22**, 27-33 (2009).
- [3] M. R. Dudek, S. Cebrat, M. Kowalczyk, P. Mackiewicz, A. Nowicka, D. Mackiewicz, M. Dudkiewicz, *Computat. Meth. Sci. Technol.* **13**, 5–12 (2007).
- [4] O. Ogasawara, Sh. Kawamoto and K. Okubo, *Comptes Rendus Biologies*, **326**, 1097-1101 (2003).
- [5] C. M. Urzúa, *Economics Letters*, **66**, 257-260 (2000).
- [6] T. R. Gulten, *Politics and the Life Sciences* **21**, 26–36 (2002).
- [7] A. Chakraborti and M. Patriarca, *Phys. Rev. Lett.* **103**, 228701 (2009).
- [8] T. Knudsen, *Am. J. Econ. Sociol.* **60**, 123–146 (2001).
- [9] J. C. Bohorquez, S. Gourley, A. R. Dixon, M. Spagat, and N. F. Johnson, *Nature* **462**, 911-914 (2009).
- [10] I. Kanter and D. A. Kessler, *Phys. Rev. Lett.* **74**, 4559 (1995).
- [11] J. F. Fontanari and L. I. Perlovsky, *Phys. Rev. E* **70**, 042901 (2004).
- [12] R. Ferrer i Cancho, *Physica A* **345**, 275-284 (2005).
- [13] K. E. Kechedzhi, O. V. Usatenko, and V. A. Yampol'skii, *Phys. Rev. E* **72**, 046138 (2005).

- [14] S. Bernhardsson, L. E. Correa da Rocha, P. Minnhagen, New J. Phys. **11**, 123015 (2009); Physica A **389** 330-341 (2010).
- [15] C. von Ferber, T. Holovatch, Yu. Holovatch, V. Palchykov, Eur. Phys. J. B **68**, 261–275 (2009).
- [16] B. Mandelbrot, in: *Communication Theory*, ed. by W. Jackson (Academic, New York, 1953), pp. 486–502.
- [17] H. de Campos and J. M. Tolman, Poetics Today **3**, 177-187 (1982).
- [18] K. Kosmidis, A. Kalampokis, P. Argyrakis, Physica A **366**, 495-502 (2006).
- [19] S. Miyazima and K. Yamamoto, Fractals, **16**, 25-32 (2008).
- [20] I.-I. Popescu et al., *Word Frequency Studies* (Mouton de Gruyter, Berlin–New York, 2009).
- [21] S. Buk, A. Rovenchak, in: *Stežkamy Frankovoho tekstu: Komunikatyvni, stylistyčni ta leksyčni vymiry romanu “Perekhresni stežky” [By the paths of Franko’s text: Communicative, stylistic, and lexical dimensions of the novel Cross-Paths]* (Lviv, Lviv University Press, 2007), pp. 138–369.
- [22] E. G. Hirsch, I. M. Casanowicz, J. Jacobs, M. Schloessinger, in: *The Jewish Encyclopedia* (Funk and Wagnalls, New York, 1901–1906), pp. 226–229; available online at <http://www.jewishencyclopedia.com>.
- [23] M. C. Clarke, *The complete concordance to Shakspeare* (New York, Wiley and Putman, 1846)
- [24] A. Kornai, *Mathematical Linguistics* (Springer, 2008).
- [25] J. P. Tuldava, *Problemy i metody kvantitativno-sistemnogo issledovanija leksiki [Problems and methods of the quantitative-systemic study of lexics]* (Tallinn, Valgus, 1987).
- [26] V. Bagnato, D. Kleppner, Phys. Rev. A **44**, 7439 (1991).
- [27] A. Posazhennikova, Rev. Mod. Phys. **78**, 1111 (2006).
- [28] E. Kelih, S. Buk, P. Grzybek, A. Rovenchak, in: *Methods of Text Analysis* (Chernivtsi: ČNU, 2009), pp. 125-132.

List of Figures

1	Typical rank–frequency distribution. The absolute frequency f is shown versus the rank r for orthographic words of <i>Perekhresni stežky</i> [<i>The Cross-Paths</i>], a Ukrainian novel by Ivan Franko. Data are obtained by the authors on the preliminary stage of compiling the frequency dictionary of the novel [21].	3
2	(Color online) The fit of the power energy spectrum to the level occupations. Blue crosses correspond to the data obtained by the authors on the basis of first 40 chapters (of total 60) from the text mentioned in the caption of Fig. 1. Solid line is the fitting curve (4) for the first 20 values of occupation numbers N_j	6
3	(Color online) The behavior of “temperature” as the size of text grows. MD — <i>Moby-Dick</i> ; PS — <i>Perekhresni stežky</i> ; So — <i>Sobor</i>). The lines correspond to the linear fits of the data represented by the respective symbols.	7
4	(Color online) The behavior of $\ln T / \ln N$ for texts from different genres. Open letters, private letters, and sermons are shown.	9

Table 1: The parameters of “energy spectrum” and “temperature” of texts

N	α	T	$\ln T / \ln N$	T/N
<i>Moby-Dick</i> (ENG)				
5942	1.97	470.4	0.708	0.0792
39363	1.60	1773.3	0.707	0.0451
66916	1.56	2639.7	0.709	0.0394
107503	1.48	3622.3	0.707	0.0337
132968	1.48	4207.3	0.707	0.0316
165746	1.48	4968.4	0.708	0.0300
191040	1.47	5476.5	0.708	0.0287
215270	1.45	5791.3	0.706	0.0269
<i>Перехресні стежки</i> (<i>The Cross-Paths</i>) (UKR)				
343	1.57	26.6	0.562	0.0774
1052	2.03	119.1	0.687	0.1132
12949	1.68	812.0	0.708	0.0627
28010	1.73	1610.1	0.721	0.0575
40811	1.72	2270.7	0.728	0.0556
54361	1.70	2964.3	0.733	0.0545
70330	1.64	3597.4	0.734	0.0512
96083	1.57	4561.4	0.734	0.0475
ꠘꠞꠟꠤ <i>Yélén`</i> (<i>The Light</i>) journal (NKO)				
429	1.42	45.2	0.629	0.1053